DOCUMENT RESUME

ED 440 509 EC 307 799

AUTHOR Thurlow, Martha; Elliott, Judy; Ysseldyke, Jim
TITLE Out-of-Level Testing: Pros and Cons.
INSTITUTION National Center on Educational Outcomes, Minneapolis, MN.
SPONS AGENCY Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE 1999-04-00
NOTE 8p.; The Center represents a collaborative effort of the University of Minnesota, the Council of Chief State School Officers, and the National Association of State Directors of Special Education.
CONTRACT H159C50004
AVAILABLE FROM NCEO Publications Office, 350 Elliott Hall, 75 E. River Road, Minneapolis, MN 55455 ($3.50). Tel: 612-624-8561; Fax: 612-624-0879; Web site: http://www.coled.umn.edu/nceo.
PUB TYPE Collected Works - Serials (022) -- Guides - Non-Classroom (055)
JOURNAL CIT NCEO Policy Directions; n9 Apr 1999
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Decision Making; *Disabilities; *Educational Assessment; Educational Policy; Elementary Secondary Education; Evaluation Methods; *Standardized Tests; *Student Evaluation; Test Use; *Testing Problems
IDENTIFIERS *Out of Level Testing

ABSTRACT

This paper addresses the controversy about use of "out-of-level" testing, the practice of assessing students (usually those with disabilities) with a lower-level version of a test. The controversy pits unintended instructional consequences against "accurately" measuring performance and avoiding student frustration. Introductory sections explain what out-of-level testing is and offer a brief history of its use. Next, arguments for out-of-level testing are offered, including avoidance of student frustration and emotional trauma; improved accuracy of measurement; and better measurement when the context of the test matches the student's instructional level. Arguments against out-of-level testing stress that assessments must be consistent with the purpose for which they are used and that out-of-level testing reflects low expectations for students and negatively affects their instruction. Next, five assumptions for out-of-level testing and objections to these assumptions are listed. Three considerations in using out-of-level testing for individual students are identified: (1) performance on grade level assessment is likely to be spuriously higher than on out-of-level assessments; (2) instructional issues need to be addressed before students are placed in out-of-level tests; and (3) unintended consequences of out-of-level testing include never reaching grade-level or passing a high stakes test. Finally, questions for decision makers to consider before using out-of-level tests are suggested. (DB)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 440 509

NCEO POLICY DIRECTIONS

NUMBER 9

APRIL 1999

OUT-OF-LEVEL TESTING: PROS AND CONS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

EC 307799

BEST COPY AVAILABLE

NCEO POLICY DIRECTIONS

Number 9 April 1999

Out-of-Level Testing: Pros and Cons

Background

Whether called "out-of-level," "off-grade-level," "functioning-level," or "instructional-level" testing, the practice of assessing students using a lower-level version of a test is controversial. The controversy pits unintended instructional consequences against "accurately" measuring performance and avoiding student frustration. The controversy also reflects beliefs about the appropriateness of delivering instruction at a student's perceived functional level rather than adapting on-grade-level instruction to the specific needs of the student. The out-of-level testing controversy is particularly pertinent to students with disabilities, who typically are functioning at lower performance levels than their peers, and who, as a result of changes in federal education laws, must participate in state and district assessments.

To explore the controversy of out-of-level testing, and assist educators and policymakers in making appropriate decisions about its use in large-scale assessments, we describe its meaning and its history, then discuss arguments for and against its use. We conclude with several important considerations and questions to ask before implementing an out-of-level testing policy or administering an out-of-level test to a student.

What is Out-of-Level Testing?

Out-of-level testing is a term used to mean that a student who is in one grade is assessed using a level of a test that was developed for students in another grade. ***Lower-level testing*** is almost universally what is meant when terms like "out-of-level," "off-grade level," and "instructional-level" are used.

The use of out-of-level testing in large-scale assessment programs has increased during the past 10 years. Generally it is presented in policy as an accommodation or modification for students with disabilities. (See Table 1 on next page for trends in the use of out-of-level testing.) State policies often warn that scores from out-of-level testing are to be interpreted with caution; usually, out-of-level tests can be used only for students with disabilities.

History of Out-of-Level Testing

Out-of-level testing first emerged in norm-referenced testing. Norm-referenced tests (NRTs) were developed with forms for different grade levels. Originally, it was intended that a child would be given the form that corresponded to that child's grade level.

Following procedures used in individualized testing, it was sometimes decided to use the same procedures for group testing—to select for a student the form that corresponded to that student's functional skill level. The decision about which form to use was based on other information about the student, such as assessed reading level, teacher judgment of instructional level, and so on.

These approaches reflect some of the same ideas as those that are

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The College of Education & Human Development

UNIVERSITY OF MINNESOTA

Table 1. Trends in the Use of Out-of-Level Testing

States Allowing Out-of-Level Testing		
1993[a]	1995[b]	1997[c]
Georgia	Connecticut Georgia Kansas North Carolina Oregon	Alaska Connecticut Georgia Maine Missouri New Hampshire New York North Dakota Vermont West Virginia

[a] *From Thurlow, Ysseldyke, & Silverstein (1993).*
[b] *From Thurlow, Scott, & Ysseldyke (1995).*
[c] *From Roeber, Bond, & Connealy (1998).*

used in individualized intelligence and achievement testing. They are also the basis for computer-adapted testing in which performance on selected test items leads to a branch of items that start with those the student can answer correctly, regardless of the difficulty levels of the items. In fact, out-of-level testing has been called the "poor man's version of computer-adapted testing." Although out-of-level testing grew out of norm-referenced testing, it soon was being applied to criterion-referenced tests (CRTs).

▶ Perceived Pros and Cons of Out-of-Level Testing

There are both pros and cons associated with out-of-level testing. They reflect different perspectives on large-scale testing and the connection between instruction and tests developed to assess the results of instruction.

▶ Arguments For Out-of-Level Testing

Individuals who argue the pro side of out-of-level testing generally cite three types of benefits: (1) avoiding undue frustration for the student, (2) improving the accuracy of measurement, and (3) better matching the student's current educational goals and instructional level. It is suggested that it is unfair for students who are not performing at grade-level to be subjected to grade-level tests.

Avoiding student frustration and emotional trauma are common arguments for out-of-level testing. Being tested at grade-level when not performing at this level is considered to be too emotionally traumatizing, and traumatization from the testing experience is thought to increase exponentially as the difference increases between the student's grade and the grade at which the student is functioning. Those in favor of out-of-level testing also argue that out-of-level testing actually is the most humane approach for students not performing well in school. Students are not forced to dwell on their errors, but rather are provided with test items to which they can respond in a reasonable manner.

Improved accuracy of measurement is also given as a reason for out-of-level testing. Psychometric support for out-of-level testing cites the over-statement of actual performance that occurs when there are many chance-level scores for students assessed at their grade level (Doscher & Bruno, 1981; Wick, 1983). This means that the performance of students looks better than it actually is when grade-level assessments are used.

Better measurement occurs when the context of the test matches the student's instructional level. It is generally recognized that the focus of out-of-level testing may not be the same as the grade-level goals. Still, out-of-level tests are said to accurately measure the student's intermediate goals on the pathway to the grade-level standards.

▶ Arguments Against Out-of-Level Testing

Individuals who argue against the use of out-of-level testing generally focus on the purpose of assessments and concerns about expectations and instruction for students. In addition, there are specific responses to some of the arguments made by those supporting the use of out-of-level testing (see Table 2).

Assessments must be consistent with the purpose for which they are being used. Although out-of-level testing may be appropriate for making instructional decisions (e.g., knowing what skills the student has now so that plans can be made about what to teach next), it is viewed as inappropriate for accountability assessments. State and

BEST COPY AVAILABLE

district assessments almost always are used for accountability purposes—to describe what students know and can do in relation to a set of standards and to evaluate how schools and programs are progressing in providing students with desired knowledge and skills. Testing at a lower grade level does not reflect the student's performance at the standard being assessed for the majority of students.

Out-of-level testing reflects low expectations for students and negatively affects their instruction. Too often, expectations for students who have not performed well in the past are below what they should be, creating a never-ending cycle of low expectations resulting in lower performance, which in turn results in even lower expectations. There are many instances of teachers being surprised by how well students performed when they were tested at grade level. There are related concerns about what happens in instruction when out-of-level approaches are used. It may be assumed that what the student is being tested on is all that the student needs to learn, with the resulting instruction focusing on lower-level standards than those toward which the student should be striving.

Assumptions of Out-of-Level Testing

There are five assumptions that test developers say should be met before out-of-level testing is considered an appropriate adaptation of testing. And, there are objections to the appropriateness of each of the assumptions (see Table 3 on next page).

Table 2. Arguments For and Against The Use of Out-of-Level Testing

Pro Arguments	Rebuttals to Pro Arguments
Avoids student frustration and emotional trauma. This is the humane approach for students not performing well in school.	Can instruction that does not address needed grade-level material be thought of as humane? Trauma will be a non-issue if instruction is consistent with the difficulty of the assessment.
Improves accuracy of measurement.	How can a test that does not address "grade-level" materials be more or less accurate than chance scores?
Better matches the student's current educational goals and instructional level.	Is it honest for a test to conform to where it is perceived a child is rather than match what we want the child to know and be able to do?

Considerations in Using Out-of-Level Testing

Three considerations derived from research are important when thinking about the use of out-of-level testing either for a system or for individual students:

Performance on grade-level assessments is likely to be spuriously higher than on out-of-level assessments. Doscher and Bruno (1981) identified this trend in a simulation of test performance, noting that "results show test scores to be overstatements of subject mastery, with larger distortions at the lower achievement levels" (p. 475). Wick (1983) confirmed this tendency using actual standardized test results from the Chicago Public Schools, where frequent occurrences of chance scores produced overstatements of actual performance. Wick produced data showing that a move from on-grade testing to functional testing resulted in lower scores, with the negative impact increasing as the students' grade-level increased.

Instructional issues need to be addressed before students are placed in out-of-level tests. Too often, assessments are seen as entities in themselves, unrelated to the instruction that is to be reflected in test performance. Discussions of out-of-level testing must return to discussions about out-of-level *teaching*. For some time, we have known that assessments are linked in varying degrees to the curricula that they reflect.[1] A disconnect between the two at any point can lead to questions about

BEST COPY AVAILABLE

Table 3. Assumptions For Out-of-Level Testing

Assumptions for Out-of-Level Testing	Objections to Assumptions
The performance of any small subgroup (such as students with disabilities) will not have a significant effect on test statistics. Although it is desirable to include a wide range of students during test development, it is not necessary that small subgroups be included.	**Reporting the results for students with disabilities is required by law.** To be able to report accurately on the performance of what may seem to be a small group of students, their inclusion during test development is critical.
The test has levels, with each successive level reflecting more difficult content on the same scale. The levels of a test typically correspond to grade levels, and the performance of students in the targeted grade, and in nearby grades, can be represented as a distribution of scores reflecting increasing difficulty. For out-of-level testing, different levels must be on the same scale, which is most easily achieved when the test has an Item Response Theory (IRT) basis.[2]	**Levels-based tests reflect different content as well as more difficult content.** Creating test levels that are on the same scale of difficulty does not adjust for the different content that may be included in different levels. Assuming that the content is the same can lead to erroneous conclusions (e.g., a grade 3 scale score used as if it were a grade 5 scale score might imply that a student has some mastery of long division when the grade 3 test does not include any items on long division).
The student is in the same scope and sequence as other students in the same grade. This means that the student is working on the same academic content as the focus of the assessment (same standards), although it may be at a much lower level. For example, out-of-level testing on a reading test can be considered for a student who is learning to read, even if at a lower level than other students in the same grade, but not for a student who is learning feeding or other self-care skills.	**The delineation of what is in the same scope and sequence is not clear.** If reading is interpreted broadly, a student who is learning to recognize letters (pre-reading skill) could be considered to be in the same scope and sequence as students who are learning to read narrative writing. Typical state and district tests do not cover such a broad range of skills. Furthermore, having broad skills does not fix the problem of different content.
Out-of-level testing is appropriate for system, not student, accountability. When the focus of the assessment is system accountability, then out-of-level testing is appropriate because it provides the best estimate of all students' skills. It is not appropriate for student accountability because deciding that a student needs a lower level test is a declaration that the student does not have mastery before the student takes the test.	**Out-of-level testing is also inappropriate for system accountability.** Because out-of-level testing does not really tap what the curriculum is for those students tested out of level, it is not appropriate to judge the system using a test that does not address what the student should know and be able to do.
Scores from out-of-level testing must be transformed to scale scores. It is inappropriate to assume that a raw score from an out-of-level test (or a percentile rank based on a raw score) can be reported along with scores from on-grade level assessments.	**Even transformed scores do not necessarily mean the same thing.** Using converted scale scores may be just as inappropriate because the out-of-level test reflects different content also.

the validity of the test performance.

When decision makers are considering the use of out-of-level testing for a particular student, their thoughts should immediately turn to the appropriateness of instruction, and its link to the assessment. Thinking only about the assessment allows one to ignore the critical element—the student's instruction. Decision makers must be able to justify the decision and at the same time be able to defend the instruction that should provide the basis for on-level test performance.

Unintended consequences of out-of-level testing include never reaching grade-level, or passing a high stakes test. The use of out-of-level testing creates unintended

BEST COPY AVAILABLE

consequences beyond simply having low expectations for a student. When the assessment for which out-of-level testing is being considered is one that is high stakes for the student—such as a graduation exam—the use of out-of-level testing essentially prevents the student from passing the high stakes assessment. Unless specific procedures are in place for immediately moving a student into a grade-level assessment when the student does well on an out-of-level assessment, the student probably will never reach grade level or pass a test that determines promotion or graduation.

Questions to Ask Before Using Out-of-Level Tests

Because there are both pros and cons to the use of out-of-level testing, it is extremely important to make good decisions about whether out-of-level testing is appropriate for an individual student, given the purpose of the assessment. Likewise, it is important for testing programs to consider the potential consequences of providing out-of-level testing as an option that may be selected for individual students. There are several questions that should be asked to help decision makers formulate good decisions.

What is the purpose of the assessment? If the purpose of the assessment is student accountability, then the use of out-of-level testing may make it essentially impossible for those students given out-of-level tests to pass the tests. This is because the lower level tests do not allow the student to demonstrate mastery at the required difficulty level. If the purpose of the assessment is system accountability, and there is no need to report on the performance of any subgroup of students that might have a significant proportion taking out-of-level assessments, then out-of-level assessments are possibly appropriate. If there is a need to be able to report on a subgroup of students likely to be put into out-of-level assessments, such as students with disabilities, then the use of out-of-level tests creates a problem for being able to report accurate scores.

Was the test designed to have different levels that are appropriately connected? Most large-scale assessments used by districts or states, unless they are norm-referenced assessments, were not designed to have different levels that correspond to different grades or groups of grades. Still, tests developed via Item Response Theory (IRT) have the potential to provide the information needed to form a common scale across disparate grade levels. Those considering out-of-level testing must make sure that a common scale is available across levels to have a psychometric justification for the use of out-of-level testing.

Are the unintended consequences of out-of-level testing appropriate? If the assessment is used to drive changes in instructional practices, then serious questions must be raised about the use of out-of-level assessments. If a test is directed to anything less than what you want the student to know and be able to do, the danger of reinforcing inappropriate instruction is considerable. On the other hand, if a test is used to determine what skills to next teach a student, then out-of-level testing may be appropriate, as long as the end goal, the standard, is still in sight.

Summary

While there are times when out-of-level testing may be appropriate, there are many times when it is not. Careful consideration of the assumptions underlying out-of-level testing, the purpose of the assessment and its characteristics, and the potential consequences of using out-of-level testing is advised for any program or individual decision-making team contemplating the use of out-of-level testing.

Resources

Simulation of Inner-City Standardized Testing Behavior: Implications for Instructional Evaluation. Doscher, M., & Bruno, J. E. (1981). American Educational Research Journal, 18 (4), 475-489.

Annual Survey of State Student Assessment Programs Fall 1997 (Vol II). Roeber, E., Bond, L., & Connealy, S. (1998). Washington, DC: Council of Chief State School Officers.

Testing Accommodations for Students with Disabilities: A Review of the Literature (Synthesis Report 4). Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

A Compilation of States' Guidelines for Accommodations in Assessments for Students with Disabilities (Synthesis Report 18). Thurlow, M. L., Scott, D. L., & Ysseldyke, J. E. (1995a). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

A Compilation of States' Guidelines for Including Students with Disabilities in Assessments (Synthesis Report 17). Thurlow, M. L., Scott, D. L., & Ysseldyke, J. E. (1995b). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Reducing Proportion of Chance Scores in Inner-City Standardized Testing Results: Impact on Average Scores. Wick, J. W. (1983). American Educational Research Journal, 20 (3), 461-463.

Footnotes

[1] Research indicating that performance on assessments is linked to curricula includes that of Shriner and Salvia (1988), who demonstrated that performance on mathematics tests varied as a function of the curriculum that was the basis for the student's instruction, and Bielinski and Davison (1998), who showed that differences in the construction of mathematics items can account for differences in the performance of males and females on the SAT.
[Shriner, J., & Salvia, J. (1988). Content validity of two tests with two math curricula over three years: Another instance of Chronic noncorrespondence. *Exceptional Children, 55*, 240-248]
[Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal, 35* (3), 455-476]

[2] Item Response Theory is one approach to constructing tests. It is based on the characteristics of individual test items. To create common scale scores, different levels of the test are administered to the same students. For example, a state with tests in grades 3 and 5 might link the two tests by administering both to a sample of grade 4 students. Raw scores on the two tests are linked to form a common scale score. In this way, for example, it is determined that a raw score of 35 on the grade 3 level test is approximately equivalent to a raw score of 20 on the grade 5 level test. Both of these are translated to a scale score of, say, 350.

Appreciation is extended to Mark Davison (Professor, University of Minnesota), for the hours he spent informing us about the psychometric justification for out-of-level testing, and to Scott Trimble (Assessment Director, Kentucky Department of Education), who provided extended comments on drafts and brought the realism of state assessments to the issues addressed here. ▲

About NCEO

The National Center on Educational Outcomes (NCEO) was established in 1990 to provide national leadership in the identification of outcomes and indicators to monitor educational results for all students, including students with disabilities. NCEO addresses the participation of students with disabilities in national and state assessments, standards-setting efforts, and graduation requirements.

The Center represents a collaborative effort of the University of Minnesota, the Council of Chief State School Officers (CCSSO), and the National Association of State Directors of Special Education (NASDSE).

The Center is supported through a Cooperative Agreement with the U.S. Department of Education, Office of Special Education Programs (H159C50004). Opinions or points of view do not necessarily represent those of the U.S. Department of Education or Offices within it.

BEST COPY AVAILABLE

NCEO Staff

Robert H. Bruininks
Michael L. Moore
Dorene L. Scott
Sandra J. Thompson
Martha L. Thurlow, associate director
James E. Ysseldyke, director

The University of Minnesota is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation.

NCEO *Policy Directions* is a series of reports that address national policy issues related to students with disabilities. This report was prepared by Martha Thurlow, Judy Elliott, and Jim Ysseldyke, with input from many individuals. It is available in alternative formats upon request.

Additional copies may be ordered for $3.50. Please write:

Publications Office
NCEO
350 Elliott Hall
75 East River Road
University of Minnesota
Minneapolis, MN 55455
Phone: 612-624-8561
Fax: 612-624-0879

Visit our Web site at http://www.coled.umn.edu/nceo/